

The exact information-based complexity of smooth convex minimization

Yoel Drori*

June 7, 2016

Abstract

We obtain a new lower bound on the information-based complexity of first-order minimization of smooth and convex functions. We show that the bound matches the worst-case performance of the recently introduced Optimized Gradient Method, thereby establishing that the bound is tight and can be realized by an efficient algorithm. The proof is based on a novel construction technique of smooth and convex functions.

Keywords Convex optimization; Complexity; Rate of convergence; Information-based complexity

1 Introduction

The problem of smooth and convex minimization plays a key role in a various range of applications, including signal and image processing, communications, machine learning, and many more. Some of the most successful approaches for solving these problems are first-order methods, i.e., algorithms that are only allowed to gain information on the objective by evaluating its value and gradient. The benefit of limiting the amount of accessible information is that these algorithms generally involve very cheap and simple computational iterations, making them suitable for tackling large scale problems. This benefit, however, comes with a price: first-order methods, in general, require considerable computational effort in order reach highly accurate solutions, making them practical when only moderate accuracy is sufficient.

As the scale of modern problems grows and finding efficient algorithms becomes increasingly important, a natural question that arises, and will be the main focus of this paper, is finding the level of accuracy that can be attained by first-order methods using a bounded computational effort. Note that there is some difficulty in answering this question

*Google Inc. (e-mail: dyoel@google.com)

that originates from the fact that the computational effort of a first-order method consists of two parts: the effort in choosing the points where the objective is to be evaluated (called the *search points*) and the effort in calculating the objective value and gradient at these points. Observing that the evaluation of the objective and its gradient often dominates the computational effort of the computation and following the theory of information-based complexity introduced in [8], we resolve this issue by measuring the computational effort of an algorithm by the number of times it evaluated the objective and its gradient, neglecting the effort required for choosing the search points.

To put these concepts in more precise terms, consider the following unconstrained problem

$$(P) \quad f^* = \min_{x \in \mathbb{R}^d} f(x),$$

where f is a smooth and convex function. A *first-order* optimization method is an iterative algorithm that approximates the solution of (P) , where it is only allowed to gain information on the objective f via a first-order oracle, \mathcal{O}_f , that is, a subroutine which given a point in \mathbb{R}^d , returns the value of the objective and its gradient at that point. In addition, since the problem of unconstrained minimization is invariant under translations, we also assume that the algorithm is provided with a reference (or starting) point $x_0 \in \text{dom}(f)$ that is often assumed to be “not too far” from an optimal solution. We call the pair (\mathcal{O}_f, x_0) a *problem instance*, and for a first-order method A we denote the approximate solution generated by algorithm when applied on this problem instance by $A(\mathcal{O}_f, x_0)$ ¹.

Within the setting considered above, a commonly used criterion for measuring the accuracy of an approximate solution is the *absolute inaccuracy* criterion, which quantifies the accuracy of an approximate solution ξ for a problem instance (\mathcal{O}_f, x_0) by the value of $f(\xi) - f^*$ (for alternative criteria see e.g., [7, Section 3.3]). Under this criterion, the *efficiency estimate* of a first-order method A over some given set of problem instances \mathcal{I} is defined as the worst-case absolute inaccuracy of A , i.e.,

$$\varepsilon(A; \mathcal{I}) := \sup_{(\mathcal{O}_f, x_0) \in \mathcal{I}} f(A(\mathcal{O}_f, x_0)) - f^*.$$

We can now put the main concept addressed in this paper in formal terms: denoting by \mathcal{A}_N the set of all first-order methods that perform at most $N \in \mathbb{N}$ calls to their first-order oracle, the *minimax risk function* [5] associated with \mathcal{I} is defined as the infimal efficiency estimate that a first-order method can attain over \mathcal{I} as a function of the computational effort N :

$$\mathcal{R}_{\mathcal{I}}(N) := \inf_{A \in \mathcal{A}_N} \varepsilon(A; \mathcal{I}).$$

Note that the classical notion of *information-based complexity* of the set \mathcal{I} can be identified as the inverse to the risk function,

$$\mathcal{C}_{\mathcal{I}}(\varepsilon) := \min\{N : \mathcal{R}_{\mathcal{I}}(N) \leq \varepsilon\},$$

¹In order to simplify the presentation, we assume A has a built-in stopping criterion that does not depend on external input.

i.e., the minimal computational effort needed by a first-order method in order to reach a given worst-case accuracy level, however, in the following we express our results using the minimax risk function as it proves to be more convenient.

Clearly, in order to establish an upper bound on the minimax risk of a class it is sufficient to find an upper bound on the efficiency estimate of a single first-order method (the main problem here being the identification of a good algorithm). On the other hand, establishing lower bounds on the minimax risk requires a more involved analysis, as the bound needs to hold for any first-order method. Several approaches appear in the literature for establishing lower-bounds, including resisting oracles [9], construction of a “worst-case” function [5, 10], and reduction to statistical problems [1, 11, 13].

Note that existing works on information-based complexity focus mainly on establishing order of magnitude bounds, where less attention is paid to absolute constants. Nevertheless, the exact minimax risk was established for several important classes of problem, some of which are detailed below.

Exact minimax risk results Consider the problem of convex quadratic minimization:

$$\mathcal{P}_{\rho,R}(\mathbb{R}^d) := \{(\mathcal{O}_f, x_0) : f(x) \equiv \frac{1}{2}x^T A x + b^T x + c, \text{ for } x \in \mathbb{R}^d, A \succeq 0, \|A\| \leq \rho, \\ \|x_0 - x_*\| \leq R \text{ for some } x_* \in X_*(f)\}.$$

It was established by Nemirovsky in [8, §2.3.B] that for $d \geq 2N + 3$ exact minimax risk associated with this class is given by

$$\mathcal{R}_{\mathcal{P}_{\rho,R}(\mathbb{R}^d)}(N) = \frac{\rho R^2}{2(2N + 1)^2}. \quad (1.1)$$

Nemirovsky also shows in [8] that this bound is attained by the Tschebyshev Methods, and in a later work, that this bound is attained by the well-known and efficient Conjugate Gradient method (this is a somewhat forgotten result, see (5.4.22) in [7], where unlike classical bounds on the Conjugate Gradient method, this bound does not depend on nontrivial spectral properties of A).

Another fundamental class of problems for which the exact minimax risk result is known is the class of non-smooth convex functions,

$$\mathcal{C}_{M,R}(\mathbb{R}^d) := \{(\mathcal{O}_f, x_0) : f \text{ is a convex function in } C_M^{0,0}(\mathbb{R}^d), \\ \|x_0 - x_*\| \leq R \text{ for some } x_* \in X_*(f)\}.$$

For this class, the exact minimax risk was recently established in [4], where it was shown that for $d \geq 2N + 1^2$

$$\mathcal{R}_{\mathcal{C}_{M,R}(\mathbb{R}^d)}(N) = \frac{MR}{\sqrt{N + 1}}. \quad (1.2)$$

²This bound can be improved to cover the case $d \geq N + 1$, see Remark 4.1.

Note that this bound is obtained by several efficient methods, including the subgradient method [10, §3.2.3] and also a family of methods recently studied in [4], which are similar to Kelley's well-known Cutting-Plane Method.

In this paper we focus on the class of smooth convex functions with Lipschitz-continuous gradient:

$$\mathcal{F}_{L,R}(\mathbb{R}^d) := \{(\mathcal{O}_f, x_0) : f \text{ is a convex function in } C_L^{1,1}(\mathbb{R}^d), \\ \|x_* - x_0\| \leq R, \text{ for some } x_* \in X_*(f)\}.$$

Since any convex quadratic function is also a smooth function with Lipschitz-continuous gradient with constant $L = \|A\|$, we have $\mathcal{P}_{L,R}(\mathbb{R}^d) \subset \mathcal{F}_{L,R}(\mathbb{R}^d)$ and thus we get that the bound (1.1) also forms a lower bound on the minimax risk of the more general class of convex functions with Lipschitz-continuous gradients. To the best of our knowledge, this bound is currently the best known lower bound on this class of problems.

Concerning an upper bound on $\mathcal{R}_{\mathcal{F}_{L,R}(\mathbb{R}^d)}(N)$, a new method, called the Optimized Gradient Method, was recently introduced in [3, 6]. It was shown (first numerically in [3], then analytically [6]) that a sequence x_0, \dots, x_N generated by this method for some $N \in \mathbb{N}$, satisfies [6, Theorem 2]:

$$f(x_N) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2\theta_N^2}, \quad \forall x_* \in X_*(f),$$

where

$$\begin{aligned} \theta_0 &= 1, \\ \theta_i &= \frac{1 + \sqrt{1 + 4\theta_{i-1}^2}}{2}, \quad i = 1, \dots, N-1, \\ \theta_N &= \frac{1 + \sqrt{1 + 8\theta_{N-1}^2}}{2}, \end{aligned} \tag{1.3}$$

which establishes the following upper bound on the minimax risk of smooth minimization:

$$\mathcal{R}_{\mathcal{F}_{L,R}(\mathbb{R}^d)}(N) \leq \frac{LR^2}{2\theta_N^2}. \tag{1.4}$$

Note that since $\sqrt{1 + 4\theta_{i-1}^2} = 2\theta_{i-1} + o(1)$, then for $i < N$ we have $\theta_i = i/2 + o(i)$ and

$$\theta_N = \frac{\sqrt{2}}{2}N + o(N),$$

hence there exists a gap of about a factor of eight between the lower bound (1.1) and the upper bound (1.4). The main goal of this paper is to close this gap by showing that the bound (1.4) is in fact tight, i.e., the inequality (1.4) can be turned into an equality.

The approach taken by this paper is motivated by the “worst-case function” proof technique introduced in [10] and the construction of smooth and convex functions developed in [14].

Overview of the paper The rest of this paper is organized as follows. We begin, in Section 2, by introducing a novel construction of smooth and convex functions that satisfy a given set of requirements on their value and gradient. We then use this construction, in Section 3, to define a function that possesses properties making it suitable for constructing lower bounds on the minimax risk function. Building on these results, in Section 4, we establish the minimax risk associated with $\mathcal{F}_{L,R}(\mathbb{R}^d)$. Finally, in Section 5 give some concluding remarks.

Notations We denote by $\mathbf{u}_i := e_{i+1}$ the canonical unit vectors with a *zero-based* index (e.g., the canonical basis vectors for \mathbb{R}^n are denoted by $\mathbf{u}_0, \dots, \mathbf{u}_{n-1}$). For a convex function f , we denote by $X_*(f)$ its set minimizers, which we assume to be nonempty.

2 A smooth convex interpolation scheme

In this section, we describe a general construction of smooth and convex functions that satisfy a set of first-order requirements. This construction can be viewed as a generalized primal form of the interpolation scheme developed in [14].

More precisely, given a finite index set I and a set of triples $\mathcal{T} = \{(x_i, g_i, f_i)\}_{i \in I}$ with $x_i \in \mathbb{R}^d$, $g_i \in \mathbb{R}^d$, and $f_i \in \mathbb{R}$ for some $d \in \mathbb{N}$, we proceed to define a convex function that has a Lipschitz-continuous gradient with constant $L > 0$ and, under certain natural conditions, is *interpolating* through the set \mathcal{T} , i.e., for all $i \in I$ the value of the function at x_i is f_i and its gradient at this point is g_i .

Definition 2.1. Let $L > 0$, let \mathcal{T} be as defined above, and denote by $w_{\mathcal{T}} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^I \rightarrow \mathbb{R}$ the following convex quadratic function:

$$w_{\mathcal{T}}(y, \nu, \alpha) := \frac{L}{2} \|y + \nu - \sum_{i \in I} \alpha_i (x_i - \frac{1}{L} g_i)\|^2 + \sum_{i \in I} \alpha_i (f_i - \frac{1}{2L} \|g_i\|^2). \quad (2.1)$$

Then for any closed convex set $C \subset \mathbb{R}^d$ such that $0 \in C$, we define the *primal interpolating function of \mathcal{T} with kernel C* , $W_{\mathcal{T}}^C(y) : \mathbb{R}^d \rightarrow \mathbb{R}$, by

$$W_{\mathcal{T}}^C(y) := \min_{\nu \in C, \alpha \in \Delta_I} \{w_{\mathcal{T}}(y, \nu, \alpha)\}, \quad (2.2)$$

where $\Delta_I := \{\alpha \in \mathbb{R}^I : \sum_{i \in I} \alpha_i = 1, \alpha_i \geq 0, \forall i \in I\}$ is an $|I|$ -dimensional unit simplex.

Note that the extra degree of freedom granted by the inclusion of the set C , although not necessary for the purpose of finding an interpolating function, will be crucial for establishing the properties of the worst-case function in the next section.

For later reference, we now state some immediate necessary and sufficient optimality conditions for the minimization problem $W_{\mathcal{T}}^C(y)$.

Lemma 1. Fix any $y \in \mathbb{R}^d$ and suppose (ν^*, α^*) is feasible for the convex optimization problem $W_{\mathcal{T}}^C(y)$. Then (ν^*, α^*) is optimal for $W_{\mathcal{T}}^C(y)$ if and only if the following conditions hold:

$$\nu^* = P_C(-y + \sum_{i \in I} \alpha_i^* (x_i - \frac{1}{L} g_i)), \quad (2.3)$$

and for any $j, k \in I$ such that $\alpha_j^* > 0$

$$\begin{aligned} & -L \langle x_j - \frac{1}{L} g_j, y + \nu^* - \sum_{i \in I} \alpha_i^* (x_i - \frac{1}{L} g_i) \rangle + (f_j - \frac{1}{2L} \|g_j\|^2) \\ & \leq -L \langle x_k - \frac{1}{L} g_k, y + \nu^* - \sum_{i \in I} \alpha_i^* (x_i - \frac{1}{L} g_i) \rangle + (f_k - \frac{1}{2L} \|g_k\|^2). \end{aligned} \quad (2.4)$$

Here P_C denotes the projection onto the set C .

Proof. Condition (2.3) follows directly from the definition of the projection function, and condition (2.4) is the first-order optimality condition for α . See for example [2]. \square

The main property of $W_{\mathcal{T}}^C(y)$ is summarized by the following theorem.

Theorem 1. The function $W_{\mathcal{T}}^C(y)$ is convex and in $C_L^{1,1}$. Furthermore, $W_{\mathcal{T}}^C(x_i) = f_i$ and $\nabla W_{\mathcal{T}}^C(x_i) = g_i$ for any $i \in I$ that satisfies

$$P_C(-\frac{1}{L} g_i) = 0, \quad (2.5)$$

and

$$\frac{1}{2L} \|g_i - g_j\|^2 \leq f_j - f_i - \langle g_i, x_j - x_i \rangle, \quad \forall j \in I. \quad (2.6)$$

Proof. Convexity. This follows from a well-known property of the infimum operator. See e.g., [12, Proposition 2.22].

Lipschitz-continuity. We first show that $W_{\mathcal{T}}^C$ is differentiable. Let $y_0 \in \mathbb{R}^d$ and suppose $(\nu^{(0)}, \alpha^{(0)})$ is an optimal solution to the optimization problem $W_{\mathcal{T}}^C(y_0)$, then it follows directly from the definition of the subdifferential and the definition of $W_{\mathcal{T}}^C$ that

$$\partial W_{\mathcal{T}}^C(y) \subseteq \partial_y w_{\mathcal{T}}(y, \nu^{(0)}, \alpha^{(0)}) = \{\nabla_y w_{\mathcal{T}}(y, \nu^{(0)}, \alpha^{(0)})\}.$$

Since $W_{\mathcal{T}}^C(y)$ is convex and defined over its entire domain, $\partial W_{\mathcal{T}}^C(y)$ is nonempty and we get that $\partial W_{\mathcal{T}}^C(y) = \{\nabla_y w_{\mathcal{T}}(y, \nu^{(0)}, \alpha^{(0)})\}$, i.e., $W_{\mathcal{T}}^C(y)$ is differentiable.

In order to show that $W_{\mathcal{T}}^C$ has Lipschitz-continuous gradient, suppose $y_1, y_2 \in \mathbb{R}^d$ and let $(\nu^{(1)}, \alpha^{(1)})$ be an optimal solution to the optimization problem $W_{\mathcal{T}}^C(y_1)$. Since $w_{\mathcal{T}}(y, \nu^{(1)}, \alpha^{(1)})$ has a Lipschitz-continuous gradient with constant L with respect to the variable y , we get

$$\begin{aligned} W_{\mathcal{T}}^C(y_2) & \leq w_{\mathcal{T}}(y_2, \nu^{(1)}, \alpha^{(1)}) \\ & \leq w_{\mathcal{T}}(y_1, \nu^{(1)}, \alpha^{(1)}) + \langle \nabla_y w_{\mathcal{T}}(y_1, \nu^{(1)}, \alpha^{(1)}), y_2 - y_1 \rangle + \frac{L}{2} \|y_1 - y_2\|^2 \\ & = W_{\mathcal{T}}^C(y_1) + \langle \nabla W_{\mathcal{T}}^C(y_1), y_2 - y_1 \rangle + \frac{L}{2} \|y_1 - y_2\|^2, \end{aligned}$$

which is desired inequality.

Interpolation property. In order to establish this part of the claim, we first show that when (2.5) and (2.6) hold, then for all $i \in I$ an optimal solution to the optimization problem $W_{\mathcal{T}}^C(x_i)$ is given by

$$\begin{aligned} \nu^* &= 0, \\ \alpha_j^* &= \begin{cases} 1, & j = i, \\ 0, & \text{otherwise,} \end{cases} \quad \forall j \in I. \end{aligned}$$

Indeed, substituting these values in the optimality conditions, we get that (2.3) follows from (2.5), since

$$P_C(-x_i + \sum_{i \in I} \alpha_i^*(x_i - \frac{1}{L}g_i)) = P_C(-x_i + (x_i - \frac{1}{L}g_i)) = P_C(-\frac{1}{L}g_i) = 0,$$

and (2.4) becomes

$$\begin{aligned} & -L \langle x_i - \frac{1}{L}g_i, x_i - (x_i - \frac{1}{L}g_i) \rangle + f_i - \frac{1}{2L} \|g_i\|^2 \leq \\ & -L \langle x_j - \frac{1}{L}g_j, x_i - (x_i - \frac{1}{L}g_i) \rangle + f_j - \frac{1}{2L} \|g_j\|^2, \end{aligned}$$

which reduces exactly to (2.6).

The value and gradient of $W_{\mathcal{T}}^C(x_i)$ can now be determined from

$$W_{\mathcal{T}}^C(x_i) = w_{\mathcal{T}}(x_i, \nu^*, \alpha^*) = \frac{L}{2} \|x_i - (x_i - \frac{1}{L}g_i)\|^2 + f_i - \frac{1}{2L} \|g_i\|^2 = f_i$$

and

$$\nabla W_{\mathcal{T}}^C(x_i) = \nabla_y w_{\mathcal{T}}(x_i, \nu^*, \alpha^*) = L(x_i - (x_i - \frac{1}{L}g_i)) = g_i.$$

□

Remark 2.1. In this paper we do not consider the strongly-convex case, however, note that the construction above can be used to generate strongly-convex functions using the well-known property of strongly convex functions which states that a function f is strongly convex with a constant μ and has Lipschitz-continuous gradient with constant L if and only if $f(x) - \frac{\mu}{2} \|x\|^2$ is convex and has Lipschitz-continuous gradient with constant $L - \mu$.

3 A worst-case function

In this section, we construct a smooth and convex function and investigate its properties. This function will form the basis to the lower complexity proof in the next section.

Definition 3.1. Let $N \in \mathbb{N}$ and $L > 0$ be fixed and let $\zeta = (\zeta_0, \dots, \zeta_{N+2}) \in \mathbb{R}^{N+3}$ be a given vector such that

$$\zeta_0 > \dots > \zeta_{N+1} > \zeta_{N+2} = 0. \quad (3.1)$$

We define $x_i^\zeta \in \mathbb{R}^{N+1}$, $g_i^\zeta \in \mathbb{R}^{N+1}$, $f_i^\zeta \in \mathbb{R}$ for $i \in \{0, \dots, N+1\}$ as follows:

$$\begin{aligned} x_i^\zeta &:= - \sum_{j=0}^{i-1} \frac{\zeta_j - \zeta_{i+1}}{\sqrt{\zeta_j - \zeta_{j+1}}} \mathbf{u}_j, \quad i = 0, \dots, N+1, \\ g_i^\zeta &:= L \sqrt{\zeta_i - \zeta_{i+1}} \mathbf{u}_i, \quad i = 0, \dots, N, \\ g_{N+1}^\zeta &:= 0, \\ f_i^\zeta &:= \frac{L}{2} (\zeta_i + \zeta_{i+1}), \quad i = 0, \dots, N, \\ f_{N+1}^\zeta &:= 0. \end{aligned} \quad (3.2)$$

In addition, taking $\mathcal{T}^\zeta := \{(x_i^\zeta, g_i^\zeta, f_i^\zeta)\}_{i \in \{0, \dots, N+1\}}$, we set

$$w_\zeta(y, \nu, \alpha) := w_{\mathcal{T}^\zeta}(y, \nu, \alpha),$$

and

$$W_\zeta(y) := W_{\mathcal{T}^\zeta}^{\mathbb{R}_+^{N+1}}(y) \left(\equiv \min_{\nu \in \mathbb{R}_+^{N+1}, \alpha \in \Delta_{N+2}} w_\zeta(y, \nu, \alpha) \right).$$

We intentionally leave the values of ζ undefined at this point, as most properties of the function W_ζ can be established without introducing additional requirements on ζ . The specific values used in the lower complexity proof will be defined towards the end of this section.

The rest of this section is dedicated to the investigation of the special properties of $W_\zeta(y)$. We start with a few technical properties.

Proposition 1. *The following identities hold:*

$$f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2 = L\zeta_{i+1}, \quad i = 0, \dots, N+1, \quad (3.3)$$

$$-x_i^\zeta + \frac{1}{L} g_i^\zeta = \sum_{j=0}^i \frac{\zeta_j - \zeta_{i+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta \in \mathbb{R}_+^{N+1}, \quad i = 0, \dots, N+1, \quad (3.4)$$

$$\langle g_i^\zeta, -x_j^\zeta + \frac{1}{L} g_j^\zeta \rangle = L \max(\zeta_i - \zeta_{j+1}, 0), \quad i, j = 0, \dots, N+1. \quad (3.5)$$

Proof. (3.3)–(3.5) follow directly from the definitions of x_i^ζ , g_i^ζ and f_i^ζ . \square

The following corollary establishes the convexity, smoothness and interpolating properties of W_ζ . In particular, it implies that $\nabla W_\zeta(x_{N+1}^\zeta) = g_{N+1}^\zeta = 0$, hence x_{N+1}^ζ is a minimizer for W_ζ , and thus $W_\zeta^* = f_{N+1}^\zeta = 0$.

Corollary 1. *The function W_ζ is convex and has a Lipschitz-continuous gradient with constant L . In addition, W_ζ is interpolating through \mathcal{T}^ζ , i.e., $W_\zeta(x_i^\zeta) = f_i^\zeta$ and $\nabla W_\zeta(x_i^\zeta) = g_i^\zeta$ for all $i \in \{0, \dots, N+1\}$.*

Proof. By Theorem 1, it is enough to verify that x_i^ζ , f_i^ζ , and g_i^ζ satisfy conditions (2.5) and (2.6) for all $i \in \{0, \dots, N+1\}$.

Condition (2.5) follows immediately from the properties of the projection operator on the nonnegative orthant. In order to establish (2.6), from the definitions (3.2) it remains to show that for any $i \neq j \in \{0, \dots, N+1\}$,

$$\frac{L}{2}((\zeta_i - \zeta_{i+1}) + (\zeta_j - \zeta_{j+1})) \leq \frac{L}{2}(\zeta_j + \zeta_{j+1}) - \frac{L}{2}(\zeta_i + \zeta_{i+1}) - \langle g_i^\zeta, x_j^\zeta - x_i^\zeta \rangle$$

which, by $\langle g_i^\zeta, x_i^\zeta \rangle = \langle g_i^\zeta, \frac{1}{L}g_i^\zeta \rangle = 0$ for $i \neq j$ reduces to

$$L(\zeta_i - \zeta_{j+1}) \leq \langle g_i^\zeta, -x_j^\zeta + \frac{1}{L}g_j^\zeta \rangle$$

and follows from (3.5). \square

The next lemma identifies an important family of optimal solutions to the convex optimization problem W_ζ .

Lemma 2. *For any $y \in \mathbb{R}^{N+1}$, there exists an optimal solution (ν^*, α^*) to the optimization problem $W_\zeta(y)$ where $\alpha^* \in \Delta_{N+2}$ is of the form*

$$\alpha^* = (\underbrace{0, \dots, 0}_{m \text{ times}}, \alpha_m^*, 1 - \alpha_m^*, \underbrace{0, \dots, 0}_{N-m \text{ times}}) \quad (3.6)$$

for some $m \in \{0, \dots, N\}$.

Proof. We proceed by showing that for any $(\nu, \alpha) \in \mathbb{R}_+^{N+1} \times \Delta_{N+2}$ that is feasible for the optimization problem $W_\zeta(y)$, if $\alpha_{i_1} > 0$ and $\alpha_{i_3} > 0$ for some $0 \leq i_1 < i_2 < i_3 \leq N+1$ then it is possible to decrease either α_{i_1} or α_{i_3} to zero while increasing α_{i_2} without affecting the objective or violating the constraints.

Let $\delta \in \mathbb{R}^{N+2}$ be defined by

$$\begin{aligned} \delta_{i_1} &= (f_{i_2}^\zeta - \frac{1}{2L}\|f_{i_2}^\zeta\|^2) - (f_{i_3}^\zeta - \frac{1}{2L}\|f_{i_3}^\zeta\|^2) = L(\zeta_{i_2+1} - \zeta_{i_3+1}), \\ \delta_{i_2} &= (f_{i_3}^\zeta - \frac{1}{2L}\|f_{i_3}^\zeta\|^2) - (f_{i_1}^\zeta - \frac{1}{2L}\|f_{i_1}^\zeta\|^2) = L(\zeta_{i_3+1} - \zeta_{i_1+1}), \\ \delta_{i_3} &= (f_{i_1}^\zeta - \frac{1}{2L}\|f_{i_1}^\zeta\|^2) - (f_{i_2}^\zeta - \frac{1}{2L}\|f_{i_2}^\zeta\|^2) = L(\zeta_{i_1+1} - \zeta_{i_2+1}), \end{aligned}$$

with $\delta_i = 0$ for $i \notin \{i_1, i_2, i_3\}$, and, in addition, set

$$\begin{aligned} t &:= \min \left(\frac{\alpha_{i_1}}{\delta_{i_1}}, \frac{\alpha_{i_3}}{\delta_{i_3}} \right), \\ \nu' &:= \nu - t \sum_{i=0}^{N+1} \delta_i (x_i^\zeta - \frac{1}{L} g_i^\zeta), \\ \alpha' &:= \alpha - t\delta, \end{aligned}$$

then clearly either $\alpha'_{i_1} = 0$ or $\alpha'_{i_3} = 0$ and it remains to show that (ν', α') is feasible for W_ζ and attains same objective value, i.e., (i) $\nu' \in \mathbb{R}_+^{N+1}$, (ii) $\alpha' \in \Delta_{N+2}$, and (iii) $w_\zeta(y, \nu', \alpha') = w_\zeta(y, \nu, \alpha)$.

(i) In order to establish that $\nu' \in \mathbb{R}_+^{N+1}$, since $\nu \in \mathbb{R}_+^{N+1}$ and $t > 0$, it is enough to show that

$$-\sum_{i=0}^{N+1} \delta_i (x_i^\zeta - \frac{1}{L} g_i^\zeta) \in \mathbb{R}_+^{N+1}.$$

From (3.4) and the definition of δ , we have

$$\begin{aligned} -\sum_{i=0}^{N+1} \delta_i (x_i^\zeta - \frac{1}{L} g_i^\zeta) &= \sum_{k=1,2,3} \delta_{i_k} \sum_{j=0}^{i_k} \frac{\zeta_j - \zeta_{i_k+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta \\ &= \sum_{j=0}^{i_1} \frac{\delta_{i_1}(\zeta_j - \zeta_{i_1+1}) + \delta_{i_2}(\zeta_j - \zeta_{i_2+1}) + \delta_{i_3}(\zeta_j - \zeta_{i_3+1})}{\zeta_j - \zeta_{j+1}} g_j^\zeta \\ &\quad + \sum_{j=i_1+1}^{i_2} \frac{\delta_{i_2}(\zeta_j - \zeta_{i_2+1}) + \delta_{i_3}(\zeta_j - \zeta_{i_3+1})}{\zeta_j - \zeta_{j+1}} g_j^\zeta + \sum_{j=i_2+1}^{i_3} \frac{\delta_{i_3}(\zeta_j - \zeta_{i_3+1})}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \end{aligned}$$

then since the vectors g_j are in \mathbb{R}_+^{N+1} , it is sufficient to verify that for all j the coefficient of g_j is nonnegative. We consider three cases: first, suppose $j \leq i_1$, then the coefficient of g_j^ζ is:

$$L(\zeta_{i_2+1} - \zeta_{i_3+1}) \frac{\zeta_j - \zeta_{i_1+1}}{\zeta_j - \zeta_{j+1}} + L(\zeta_{i_3+1} - \zeta_{i_1+1}) \frac{\zeta_j - \zeta_{i_2+1}}{\zeta_j - \zeta_{j+1}} + L(\zeta_{i_1+1} - \zeta_{i_2+1}) \frac{\zeta_j - \zeta_{i_3+1}}{\zeta_j - \zeta_{j+1}} = 0.$$

Next, suppose $i_1 < j \leq i_2$, then the coefficient of g_j^ζ is

$$L(\zeta_{i_3+1} - \zeta_{i_1+1}) \frac{\zeta_j - \zeta_{i_2+1}}{\zeta_j - \zeta_{j+1}} + L(\zeta_{i_1+1} - \zeta_{i_2+1}) \frac{\zeta_j - \zeta_{i_3+1}}{\zeta_j - \zeta_{j+1}} = -L(\zeta_{i_2+1} - \zeta_{i_3+1}) \frac{\zeta_j - \zeta_{i_1+1}}{\zeta_j - \zeta_{j+1}},$$

which is positive since $\{\zeta_i\}$ is strictly decreasing. Lastly, for the case $i_2 < j \leq i_3$, the coefficient of g_j^ζ is

$$L(\zeta_{i_1+1} - \zeta_{i_2+1}) \frac{\zeta_j - \zeta_{i_3+1}}{\zeta_j - \zeta_{j+1}},$$

which is again positive from the monotonicity of ζ .

(ii) $\alpha' \in \Delta_{N+2}$. Firstly, the nonnegativity of α' follows from the choice of t and since $\delta_{i_1}, \delta_{i_3} > 0$ while $\delta_{i_2} < 0$. Secondly, from the definition of α' and

$$\sum_{i=0}^{N+1} \delta_i = \delta_{i_1} + \delta_{i_2} + \delta_{i_3} = 0,$$

and we get that $\sum_{i=0}^{N+1} \alpha'_i = 1$,

(iii) $w_\zeta(y, \nu', \alpha') = w_\zeta(y, \nu, \alpha)$. By the definition of w_ζ it is enough to show that

$$\begin{aligned} y + \nu - \sum_{i=0}^{N+1} \alpha_i \left(x_i^\zeta - \frac{1}{L} g_i^\zeta \right) &= y + \nu' - \sum_{i=0}^{N+1} \alpha'_i \left(x_i^\zeta - \frac{1}{L} g_i^\zeta \right) \\ \text{and } \sum_{i=0}^{N+1} \alpha_i \left(f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2 \right) &= \sum_{i=0}^{N+1} \alpha'_i \left(f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2 \right). \end{aligned}$$

Indeed, the first equality follows from

$$\nu - \nu' = t \sum_{i=0}^{N+1} \delta_i \left(x_i^\zeta - \frac{1}{L} g_i^\zeta \right) = \sum_{i=0}^{N+1} (\alpha_i - \alpha'_i) \left(x_i^\zeta - \frac{1}{L} g_i^\zeta \right),$$

and the second equality follows from

$$\begin{aligned} \sum_{i=0}^{N+1} (\alpha_i - \alpha'_i) \left(f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2 \right) &= \sum_{i=0}^{N+1} t \delta_i \left(f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2 \right) \\ &= t \left(\left(f_{i_2}^\zeta - \frac{1}{2L} \|f_{i_2}^\zeta\|^2 \right) - \left(f_{i_3}^\zeta - \frac{1}{2L} \|f_{i_3}^\zeta\|^2 \right) \right) \left(f_{i_1}^\zeta - \frac{1}{2L} \|f_{i_1}^\zeta\|^2 \right) \\ &\quad + t \left(\left(f_{i_3}^\zeta - \frac{1}{2L} \|f_{i_3}^\zeta\|^2 \right) - \left(f_{i_1}^\zeta - \frac{1}{2L} \|f_{i_1}^\zeta\|^2 \right) \right) \left(f_{i_2}^\zeta - \frac{1}{2L} \|f_{i_2}^\zeta\|^2 \right) \\ &\quad + t \left(\left(f_{i_1}^\zeta - \frac{1}{2L} \|f_{i_1}^\zeta\|^2 \right) - \left(f_{i_2}^\zeta - \frac{1}{2L} \|f_{i_2}^\zeta\|^2 \right) \right) \left(f_{i_3}^\zeta - \frac{1}{2L} \|f_{i_3}^\zeta\|^2 \right) = 0. \end{aligned}$$

Repeatedly applying the procedure described above on an optimal solution of W_ζ , (ν^*, α^*) , we get that the distance between the first and last positive elements of α_* can always be decreased as long as it is greater than one, which completes the proof. \square

Note that the previous result allows for an efficient numerical solution of $W_\zeta(y)$ via $N+1$ independent one-dimensional convex problems instead of a single $N+1$ -dimensional problem. The following theorem shows that in some cases it is possible to further reduce the number of subproblems that need to be solved in order to find a solution to $W_\zeta(y)$. More importantly, the following property turns out to be fundamental and will be later used to establish the main properties of W_ζ .

Theorem 2. Let $\bar{y} = (\bar{y}_0, \dots, \bar{y}_N)$ be a vector in \mathbb{R}^{N+1} . If $\bar{y}_k \geq 0$ for some $k \in \{0, \dots, N\}$, then $W_\zeta(\bar{y})$ has an optimal solution (ν^*, α^*) , where α^* is of the form (3.6) and, in addition, satisfies $\alpha_{k+1}^* = \dots = \alpha_{N+1}^* = 0$.

Proof. Let (ν^*, α^*) be an optimal solution to $W_\zeta(\bar{y})$ where α^* is of the form (3.6) for some $0 \leq m \leq N$. Further assume that α^* is chosen so that the value of m is minimal among all such optimal solutions to $W_\zeta(\bar{y})$.

In order to simplify the notations in the following, we denote

$$v := \alpha_m^* (-x_m^\zeta + \frac{1}{L} g_m^\zeta) + (1 - \alpha_m^*) (-x_{m+1}^\zeta + \frac{1}{L} g_{m+1}^\zeta).$$

Now, suppose that α^* is of the form (3.6) with $m \geq k$ and $\alpha_m^* < 1$, then from $\alpha_m^* < 1$, (3.5), and (3.1) we have

$$\langle g_k^\zeta, v \rangle > L(\zeta_k - \zeta_{k+1}). \quad (3.7)$$

In addition, the closed-form solution for ν^* (2.3) implies that

$$\nabla_\nu w_\zeta(\bar{y}, \nu^*, \alpha^*) = \bar{y} + \nu^* + v \in \mathbb{R}_+^{N+1}, \quad (3.8)$$

and, from the condition on \bar{y}_k , we have

$$\langle g_k^\zeta, \bar{y} + \nu^* \rangle = L\sqrt{\zeta_k - \zeta_{k+1}}(\bar{y}_k + \nu_k^*) \geq 0. \quad (3.9)$$

Combining these results we get

$$\begin{aligned} \frac{\partial w_\zeta(\bar{y}, \nu^*, \alpha^*)}{\partial \alpha_k} &= \langle -x_k^\zeta + \frac{1}{L} g_k^\zeta, \bar{y} + \nu^* + v \rangle + f_k^\zeta - \frac{1}{2L} \|g_k^\zeta\|^2 \\ &= \left\langle \sum_{j=0}^{k-1} \frac{\zeta_j - \zeta_{k+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + \langle g_k^\zeta, \bar{y} + \nu^* + v \rangle + L\zeta_{k+1} \\ &= \left\langle \sum_{j=0}^{k-1} \frac{\zeta_j - \zeta_{k+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + \langle g_k^\zeta, \bar{y} + \nu^* + v \rangle + L \frac{\zeta_{k+1} - \zeta_{m+2}}{\zeta_k - \zeta_{k+1}} (\zeta_k - \zeta_{k+1}) + L\zeta_{m+2} \\ &< \left\langle \sum_{j=0}^{k-1} \frac{\zeta_j - \zeta_{k+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + \langle g_k^\zeta, \bar{y} + \nu^* + v \rangle + \frac{\zeta_{k+1} - \zeta_{m+2}}{\zeta_k - \zeta_{k+1}} \langle g_k^\zeta, v \rangle + L\zeta_{m+2} \\ &\leq \left\langle \sum_{j=0}^{k-1} \frac{\zeta_j - \zeta_{k+1}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + \frac{\zeta_k - \zeta_{m+2}}{\zeta_k - \zeta_{k+1}} \langle g_k^\zeta, \bar{y} + \nu^* \rangle + \frac{\zeta_k - \zeta_{m+2}}{\zeta_k - \zeta_{k+1}} \langle g_k^\zeta, v \rangle + L\zeta_{m+2} \\ &\leq \left\langle \sum_{j=0}^k \frac{\zeta_j - \zeta_{m+2}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + L\zeta_{m+2} \\ &\leq \left\langle \sum_{j=0}^{m+1} \frac{\zeta_j - \zeta_{m+2}}{\zeta_j - \zeta_{j+1}} g_j^\zeta, \bar{y} + \nu^* + v \right\rangle + f_{m+1}^\zeta - \frac{1}{2L} \|g_{m+1}^\zeta\|^2 \\ &= \frac{\partial w_\zeta(\bar{y}, \nu^*, \alpha^*)}{\partial \alpha_{m+1}}, \end{aligned}$$

where the strict inequality follows from (3.7) and the following inequalities follow from (3.9), (3.1), and (3.8), respectively. By the first-order optimality conditions (2.4) it then follows that $\alpha_{m+1}^* = 0$, i.e., $\alpha_m = 1$, reaching a contradiction. We therefore conclude that either $m < k$ or $\alpha_m^* = 1$.

We are left with the following three cases:

- $m < k$: Here α^* is in the claimed form.
- $m = k = 0$ and $\alpha_0^* = 1$: Again, α^* is in the required form.
- $m > 0$, $m \geq k$, and $\alpha_m^* = 1$: This case is not consistent with the minimality of m since it is possible to represent α^* using a smaller value of m by taking $m' = m - 1$ with $\alpha_{m-1}^* = 0$.

We conclude that for all valid cases α^* is in the desired form. \square

The following two corollaries will be the main building blocks of the proof in the next section.

Corollary 2. *Suppose $\bar{y} = (\bar{y}_0, \dots, \bar{y}_N)$ is a vector in \mathbb{R}^{N+1} such that $\bar{y}_k \geq 0$ for some $k \in \{0, \dots, N\}$, then*

$$W_\zeta(\bar{y}) - W_\zeta^* \geq f_k^\zeta$$

Proof. Denote

$$\mathbb{R}^{k,N+1} := \{(y_0, \dots, y_N) \in \mathbb{R}^{N+1} : y_k = \dots = y_N = 0\},$$

and consider the case $\bar{y} \in \mathbb{R}^{k,N+1}$. Since for all $v \in \mathbb{R}^{k,N+1}$ we have $\langle \nabla W_\zeta(x_k^\zeta), v \rangle = \langle g_k^\zeta, v \rangle = L\sqrt{\zeta_i - \zeta_{i+1}} \langle e_{k+1}, v \rangle = 0$, then clearly x_k^ζ minimizes $W_\zeta(\bar{y})$ over $\mathbb{R}^{k,N+1}$, and we get that $W_\zeta(\bar{y}) - W_\zeta^* = W_\zeta(\bar{y}) \geq W_\zeta(x_k^\zeta) = f_k^\zeta$.

For the general case $\bar{y} \in \mathbb{R}^{N+1}$, let (ν^*, α^*) be an optimal solution to $W_\zeta(\bar{y})$ such that $\alpha_{k+1}^* = \dots = \alpha_{N+1}^* = 0$ (by Theorem 2 such a solution exists), and denote $\hat{y} = (\bar{y}_0, \dots, \bar{y}_{k-1}, 0, \dots, 0)$, $\hat{\nu}^* = (\nu_0^*, \dots, \nu_{k-1}^*, 0, \dots, 0)$. We get

$$\begin{aligned} W_\zeta(\bar{y}) - W_\zeta^* &= W_\zeta(\bar{y}) = w_\zeta(\bar{y}, \nu^*, \alpha^*) \\ &= \frac{L}{2} \|\bar{y} + \nu^* - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta)\|^2 + \sum_{i=0}^k \alpha_i^* (f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2) \\ &= \frac{L}{2} \|\hat{y} + \hat{\nu}^* + \sum_{i=k}^N (\bar{y}_i + \nu_i^*) \mathbf{u}_i - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta)\|^2 + \sum_{i=0}^k \alpha_i^* (f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2) \\ &= \frac{L}{2} \|\hat{y} + \hat{\nu}^* + (\bar{y}_k + \nu_k^*) \mathbf{u}_k - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta)\|^2 + \sum_{i=k+1}^N (\bar{y}_i + \nu_i^*)^2 + \sum_{i=0}^k \alpha_i^* (f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2) \\ &\geq \frac{L}{2} \|\hat{y} + \hat{\nu}^* - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta)\|^2 + \sum_{i=k+1}^N (\bar{y}_i + \nu_i^*)^2 + \sum_{i=0}^k \alpha_i^* (f_i^\zeta - \frac{1}{2L} \|g_i^\zeta\|^2) \\ &\geq w_\zeta(\hat{y}, \hat{\nu}^*, \alpha^*) \geq W_\zeta(\hat{y}) \geq f_k^\zeta, \end{aligned}$$

where the first inequality follows from $\bar{y}_k + \nu_k^* \geq 0$ and (3.4). \square

Corollary 3. Suppose $\bar{y} = (\bar{y}_0, \dots, \bar{y}_N)$ is a vector such that $\bar{y}_k \geq 0$ and $\bar{y}_n = 0$ for some $0 \leq k < n \leq N$, then

$$\frac{\partial}{\partial y_n} W_\zeta(\bar{y}) = 0.$$

Proof. Let (ν^*, α^*) be an optimal solution of $W_\zeta(\bar{y})$ in the form guaranteed by Theorem 2, then

$$\nabla W_\zeta(\bar{y}) = L \left(\bar{y} + \nu^* - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta) \right),$$

and we get

$$\frac{\partial}{\partial y_n} W_\zeta(\bar{y}) = L \langle \bar{y} + \nu^* - \sum_{i=0}^k \alpha_i^* (x_i^\zeta - \frac{1}{L} g_i^\zeta), \mathbf{u}_n \rangle = 0,$$

where the last equality follows from $\bar{y}_n = 0$, (2.3), and since $\langle x_i^\zeta - \frac{1}{L} g_i^\zeta, \mathbf{u}_n \rangle = 0$ for any $i < n$, \square

Remark 3.1. As an immediate result of Corollary 3, it follows that

$$y \in \mathbb{R}^{k, N+1} \Rightarrow \nabla W_\zeta(y) \in \mathbb{R}^{k+1, N+1}, \quad k = 0, \dots, N.$$

Using a simple inductive argument, we get that for any sequence $x_0, \dots, x_N \in \mathbb{R}^{N+1}$ starting at $x_0 = 0$ and satisfying (c.f., [10, Assumption 2.1.4])

$$x_k \in \text{span}\{\nabla W_\zeta(x_0), \dots, \nabla W_\zeta(x_{k-1})\}, \quad k = 1, \dots, N, \quad (3.10)$$

we have $x_k \in \mathbb{R}^{k, N+1}$, $k = 0, \dots, N$, and therefore, by Corollary 2, we get that the inequality

$$W_\zeta(x_k) - W_\zeta^* \geq f_k^\zeta, \quad k = 0, \dots, N \quad (3.11)$$

holds for any such sequence. We conclude that the worst-case absolute inaccuracy of any first-order method which generates sequences that satisfy (3.10) cannot be lower than the RHS of (3.11), i.e., this inequality forms a lower-bound on the efficiency estimate of such methods. As noted by Nesterov [10, Page 59], this limitation on the structure of the first-order methods for which this bound is applicable can be avoided by some additional reasoning, however, this requires introducing some additional assumptions, e.g., that the function W_ζ can be embedded in a vector space whose dimension, d , is at least $d \geq 2N + 1$. By taking advantage of the properties derived above, we show in the next section that the lower bound (3.11) applies for sequences generated by any first-order method for the case $d \geq N + 1$.

At this stage, the exact value of ζ has not yet been defined. In order to find an instance of ζ which yields the best lower complexity bounds on first-order methods, one needs to maximize f_N^ζ while keeping $\|x_{N+1}^\zeta\|$ bounded. Solving this problem, we obtain the following solution instance (we skip the optimality proof, as it follows from the tightness of the resulting bound):

Definition 3.2. Suppose $R > 0$ and $N \in \mathbb{N}$ are given. We define $\zeta^* = (\zeta_0^*, \dots, \zeta_{N+2}^*)$ by

$$\begin{aligned}\zeta_i^* &= \frac{2\theta_i}{2\theta_i - 1} \zeta_{i+1}^*, \quad i = 0, \dots, N-1, \\ \zeta_N^* &= \frac{\theta_N}{\theta_N - 1} \zeta_{N+1}^*, \\ \zeta_{N+1}^* &= \frac{\theta_N - 1}{\theta_N^2(2\theta_N - 1)} R^2, \\ \zeta_{N+2}^* &= 0,\end{aligned}\tag{3.12}$$

where θ_i is defined as in (1.3).

Lemma 3. Consider (3.2), taking $\zeta = \zeta^*$, then

$$\|x_{N+1}^{\zeta^*}\| = R \quad \text{and} \quad f_N^{\zeta^*} = \frac{LR^2}{2\theta_N^2}.$$

Proof. Using the definition of ζ_j^* and the relation

$$\theta_{i-1}^2 = \theta_i(\theta_i - 1), \quad i = 1, \dots, N-1,$$

for $j = 0, \dots, N-1$ we have

$$\begin{aligned}\zeta_j^* \theta_j &= \zeta_N^* \theta_j \prod_{i=j}^{N-1} \frac{2\theta_i}{2\theta_i - 1} \\ &= \zeta_N^* \frac{2\theta_j^2}{2\theta_j - 1} \prod_{i=j+1}^{N-1} \frac{2\theta_i}{2\theta_i - 1} = \zeta_N^* \frac{2\theta_{j+1} - 2}{2\theta_j - 1} \theta_{j+1} \prod_{i=j+1}^{N-1} \frac{2\theta_i}{2\theta_i - 1} \\ &= \zeta_N^* \frac{2\theta_{j+1} - 2}{2\theta_j - 1} \frac{2\theta_{j+1}^2}{2\theta_{j+1} - 1} \prod_{i=j+2}^{N-1} \frac{2\theta_i}{2\theta_i - 1} = \zeta_N^* \frac{2\theta_{j+1} - 2}{2\theta_j - 1} \frac{2\theta_{j+2} - 2}{2\theta_{j+1} - 1} \theta_{j+2} \prod_{i=j+2}^{N-1} \frac{2\theta_i}{2\theta_i - 1} \\ &= \dots = \zeta_N^* \prod_{i=j}^{N-2} \frac{2\theta_{i+1} - 2}{2\theta_i - 1} \cdot \theta_{N-1} \frac{2\theta_{N-1}}{2\theta_{N-1} - 1},\end{aligned}$$

then, using $\theta_{N-1}^2 = \frac{1}{2}\theta_N(\theta_N - 1)$, we reach

$$\zeta_j^* \theta_j = \frac{\zeta_N^* \theta_N}{2} \prod_{i=j}^{N-1} \frac{2\theta_{i+1} - 2}{2\theta_i - 1}, \quad j = 0, \dots, N-1.\tag{3.13}$$

Next, from the definition of $x_{N+1}^{\zeta^*}$, (3.12), and (3.13) we get

$$\begin{aligned}
\|x_{N+1}^{\zeta^*}\|^2 &= \sum_{j=0}^N \frac{(\zeta_j^*)^2}{\zeta_j^* - \zeta_{j+1}^*} = \sum_{j=0}^{N-1} \frac{\zeta_j^*}{1 - \frac{2\theta_{j+1}-2}{2\theta_j}} + \frac{\zeta_N^*}{1 - \frac{\theta_N-1}{\theta_N}} = \sum_{j=0}^{N-1} 2\zeta_j^* \theta_j + \zeta_N^* \theta_N \\
&= \sum_{j=0}^{N-1} \zeta_N^* \theta_N \prod_{i=j}^{N-1} \frac{2\theta_{i+1}-2}{2\theta_i-1} + \zeta_N^* \theta_N \\
&= \theta_N \zeta_N^* \left(\frac{2\theta_N-2}{2\theta_{N-1}-1} \left(\frac{2\theta_{N-1}-2}{2\theta_{N-2}-1} \left(\cdots \left(\frac{2\theta_2-2}{2\theta_1-1} \left(\frac{2\theta_1-2}{2\theta_0-1} + 1 \right) + 1 \right) \cdots \right) + 1 \right) + 1 \right),
\end{aligned}$$

which from $2\theta_0 - 1 = 1$ turns out to be a telescopic product and reduces to

$$\|x_{N+1}^{\zeta^*}\|^2 = \theta_N \zeta_N^* (2\theta_N - 2) = \frac{\theta_N^2 (2\theta_N - 1)}{\theta_N - 1} \zeta_{N+1}^* = R^2.$$

For the second part of the claim, we have

$$\begin{aligned}
f_N^{\zeta^*} &= \frac{L}{2} (\zeta_N^* + \zeta_{N+1}^*) \\
&= \frac{L}{2} \left(\frac{\theta_N}{\theta_N - 1} \frac{\theta_N - 1}{\theta_N^2 (2\theta_N - 1)} R^2 + \frac{\theta_N - 1}{\theta_N^2 (2\theta_N - 1)} R^2 \right) \\
&= \frac{LR^2}{2} \left(\frac{\theta_N}{\theta_N^2 (2\theta_N - 1)} + \frac{\theta_N - 1}{\theta_N^2 (2\theta_N - 1)} \right) \\
&= \frac{LR^2}{2\theta_N^2},
\end{aligned}$$

which establishes the claim. \square

4 The main result

We now build upon the results of the previous sections to derive the exact minimax risk associated with smooth and convex minimization.

We start by introducing a family of functions that will be used as a basis for proving the main result.

Definition 4.1. Let $N, d \in \mathbb{N}$ be such that $N \leq d$. For an orthonormal set $\{v_0, \dots, v_N\} \subset \mathbb{R}^d$, let $W_{v_0, \dots, v_N} : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined by

$$W_{v_0, \dots, v_N}(z) := W_{\zeta^*}(\langle z, v_0 \rangle, \dots, \langle z, v_N \rangle).$$

Lemma 4. Let $N \leq d$ and suppose $\{v_0, \dots, v_N\}$ and $\{u_0, \dots, u_N\}$ are two orthonormal sets of vectors in \mathbb{R}^d with $v_0 = u_0, \dots, v_k = u_k$ for some $k \leq N$, then for any $z \in \mathbb{R}^d$ such that $z \in \text{span}\{v_0, \dots, v_k\}$ and $\langle z, v_k \rangle \geq 0$ we have

$$W_{v_0, \dots, v_N}(z) = W_{u_0, \dots, u_N}(z) \quad \text{and} \quad \nabla W_{v_0, \dots, v_N}(z) = \nabla W_{u_0, \dots, u_N}(z).$$

Proof. Since $\langle z, u_{k+1} \rangle = \langle z, v_{k+1} \rangle = \dots = \langle z, u_N \rangle = \langle z, v_N \rangle = 0$, we get

$$\begin{aligned} W_{v_0, \dots, v_N}(z) &= W_{\zeta^*}(\langle z, v_0 \rangle, \dots, \langle z, v_k \rangle, 0, \dots, 0) \\ &= W_{\zeta^*}(\langle z, u_0 \rangle, \dots, \langle z, u_k \rangle, 0, \dots, 0) = W_{u_0, \dots, u_N}(z), \end{aligned}$$

which establishes the first part of the claim. For the second part of the claim, we have

$$\frac{\partial}{\partial v_n} W_{v_0, \dots, v_N}(z) = \frac{\partial}{\partial u_n} W_{u_0, \dots, u_N}(z),$$

which is immediate for $n \leq k$ and follows from Corollary 3 for $n > k$. \square

We are now ready to state the main result of this paper: a lower bound on the minimax risk associated with $\mathcal{F}_{L,R}(\mathbb{R}^d)$.

Theorem 3. *Let $L, R > 0$ and $N \in \mathbb{N}$, then for any first-order method A that performs at most N calls to its first-order oracle and any $d \geq N + 1$ there exists a convex function $w_A \in C_L^{1,1}(\mathbb{R}^d)$ and $x_0 \in \mathbb{R}^d$ such that*

$$\begin{aligned} \|x_* - x_0\| &\leq R, \quad \text{for some } x_* \in X_*(w_A), \\ \text{and } w_A(A(\mathcal{O}_{w_A}, x_0)) - w_A^* &\geq \frac{LR^2}{2\theta_N^2}, \end{aligned} \tag{4.1}$$

where θ_N is as defined in (1.3).

Proof. Since a first-order method can only gain information on the objective through the first-order information at the selected search points, it cannot distinguish between functions that have identical first-order information at these points, and will execute precisely in the same way on such functions. We can therefore maintain during the run of the method a set of functions that have identical first-order information at the points selected so far, postponing the choice of a specific objective until the method has completed its run.

We set $x_0 = 0$ as the reference point given to algorithm A . Now, suppose $\xi_k \in \mathbb{R}^d$ is the $(k+1)$ th search point chosen by algorithm A , $0 \leq k \leq N-1$. At this stage, we take a vector $\bar{v}_k \in \mathbb{R}^d$ that satisfies the following conditions:

- \bar{v}_k is a unit vector orthogonal to $\{\bar{v}_0, \dots, \bar{v}_{k-1}\}$,
- $\xi_k \in \text{span}\{\bar{v}_0, \dots, \bar{v}_k\}$,
- $\langle \xi_k, \bar{v}_k \rangle \geq 0$,

and set

$$\Omega_k := \{W_{v_0, \dots, v_N} : v_0, \dots, v_N \text{ are orthonormal, } v_0 = \bar{v}_0, \dots, v_k = \bar{v}_k\}.$$

By Lemma 4, all functions in Ω_k share the same first-order information at ξ_k (and also at ξ_0, \dots, ξ_{k-1} , since $\Omega_k \subset \Omega_{k-1} \subset \dots \subset \Omega_0$), consequently, algorithm A cannot differentiate

between these functions, and we can postpone the specific choice of function for a later stage.

After invoking the oracle N times, suppose algorithm A has selected $\xi_N \in \mathbb{R}^d$ as its approximate solution. Let \bar{v}_N be a unit vector orthogonal to $\{\bar{v}_0, \dots, \bar{v}_{N-1}\}$ such that $\langle \xi_N, \bar{v}_N \rangle \geq 0$, and set

$$w_A := W_{\bar{v}_0, \dots, \bar{v}_N},$$

then by the construction above, algorithm A , when applied on w_A with the reference point x_0 , will generate the search points ξ_0, \dots, ξ_{N-1} and it return as its output the vector ξ_N .

To complete the proof, note that w_A is of the form $W_{\zeta^*}(Qy)$ for some matrix $Q \in \mathbb{R}^{N+1 \times d}$ with orthonormal rows, hence w_A shares the convexity, smoothness, and Lipschitz constant of W_{ζ^*} . In addition, since $x_{N+1}^{\zeta^*}$ is a minimizer of W_{ζ^*} , we get that $Q^T x_{N+1}^{\zeta^*} \in X_*(w_A)$. Finally, from Lemma 3 and Corollary 2 we have

$$\begin{aligned} \|Q^T x_{N+1}^{\zeta^*} - x_0\| &= \|x_{N+1}^{\zeta^*}\| = R, \\ \text{and } w_A(\xi_N) - w_A^* &= w_A(\xi_N) \geq f_N^* = \frac{LR^2}{2\theta_N^2}, \end{aligned}$$

hence w_A satisfies the required properties. \square

Combining the previous theorem and the worst-case performance of the Optimized Gradient Method (1.4), we obtain the exact minimax risk associated with the minimization of smooth and convex functions:

Corollary 4. *Let $L > 0$, $R > 0$, and $N \in \mathbb{N}$, then for any $d \geq N + 1$,*

$$\mathcal{R}_{\mathcal{F}_{L,R}(\mathbb{R}^d)}(N) = \frac{LR^2}{2\theta_N^2},$$

where θ_N is as defined in (1.3).

Remark 4.1. Note that the proof of Theorem 3 can be readily applied on any function satisfying Corollaries 2 and 3 (with the obvious adjustments). One important such function is the function \bar{f}_N defined in [4, Appendix A] for establishing the minimax risk for the non-smooth case: since this function clearly satisfies the required properties, it then follows by Theorem 3 that the identity (1.2) holds for $d \geq N + 1$ (compared to $d \geq 2N + 1$ which follows by the current proof).

5 Discussion and future work

We presented a novel interpolation scheme for constructing smooth and convex functions and used it to define a function that is in some sense the worst possible for first-order methods. Taking advantage of the special properties of the resulting function, we established the exact minimax risk associated with the class of smooth and convex minimization problems.

A similar approach might be applicable for establishing bounds on additional problem classes, such as classes of problems where the approximation accuracy is measured by an alternative criteria (e.g., $\|\nabla f(x_N)\|$) and cases where the relation between x_0 and x_* is determined in an alternative way (e.g., when $\|x_0 - x_*\|_p \leq R$ for some $p \neq 2$). We leave the investigation of these problems for future work.

We conclude with the observation that, as in the cases of non-smooth minimization and convex quadratic minimization, an optimal method for smooth minimization might not be unique. Finding optimal methods with additional properties might prove to be an interesting research direction.

References

- [1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory*, 5(58):3235–3249, 2012.
- [2] D. P. Bertsekas. *Nonlinear programming*. Athena scientific, 1999.
- [3] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program. Ser. A*, 145:451–482, 2014.
- [4] Y. Drori and M. Teboulle. An optimal variant of kelley’s cutting-plane method. *Math. Program. Ser. A*, pages 1–31, 2016.
- [5] C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *J. Complexity*, 31(1):1 – 14, 2015.
- [6] D. Kim and J. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program. Ser. A*, pages 1–27, 2015.
- [7] A. Nemirovski. Optimization II: Numerical methods for nonlinear continuous optimization. *Lecture notes*, http://www2.isye.gatech.edu/~nemirovs/Lect_OptII.pdf, 1999.
- [8] A. Nemirovsky. Information-based complexity of linear operator equations. *J. Complexity*, 8(2):153–175, 1992.
- [9] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.
- [10] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Applied optimization. Kluwer Academic Publishers, 2004.

- [11] M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *Information Theory, IEEE Transactions on*, 57(10):7036–7056, 2011.
- [12] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [13] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.
- [14] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program. Ser. A*, pages 1–39, 2016.